PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : H04N 7/15, 7/26 — A1

(11) International Publication Number: WO 97/43856

(43) International Publication Date: 20 November 1997 (20.11.97)

(21) International Application Number: PCT/AU97/00297

(22) International Filing Date: 15 May 1997 (15.05.97)

(30) Priority Data:
PN 9889 16 May 1996 (16.05.96) AU

(71) Applicant *(for all designated States except US)*: UNISEARCH LIMITED [AU/AU]; 221-227 Anzac Parade, Kensington, NSW 2033 (AU).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: FRATER, Michael, R. [AU/AU]; 26 Garrad Circuit, Charnwood, ACT 2615 (AU). ARNOLD, John, Frederick [AU/AU]; 83 Summerville Crescent, Florey, ACT 2615 (AU).

(74) Agent: F.B. RICE & CO.; 28A Montague Street, Balmain, NSW 2041 (AU).

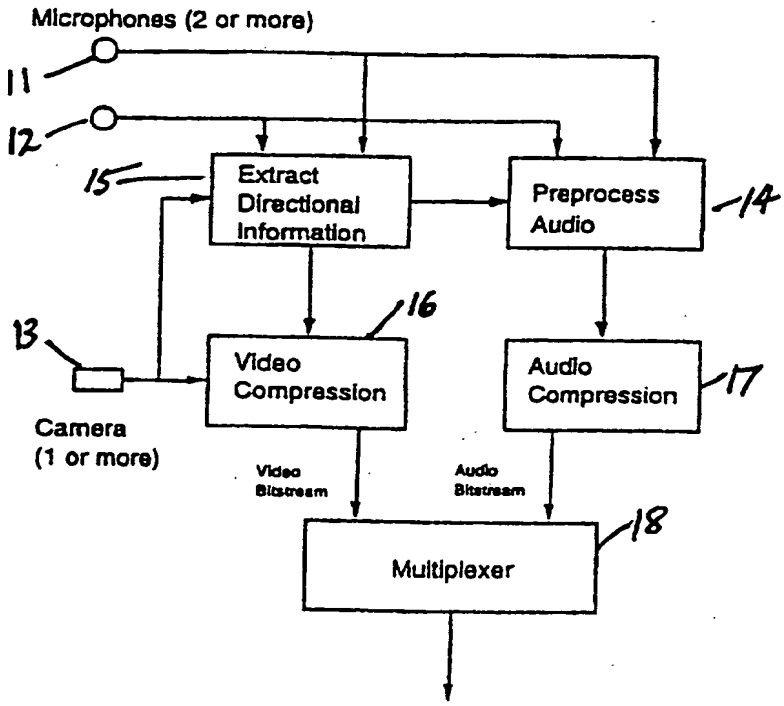
(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, HU, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ARIPO patent (GH, KE, LS, MW, SD, SZ, UG), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: COMPRESSION AND CODING OF AUDIO-VISUAL SERVICES

(57) Abstract

A new method for improving audio-visual quality is proposed which uses directional information obtained from sound (e.g. by the use of two or more microphones) to improve the quality of video services by using this directional information to identify important parts of the picture, and transmitting these parts at higher quality (e.g. by adjusting quantisation parameters used in the compression process) and by using directional information obtained from sound and from the video (e.g. by picture segmentation techniques) to improve the quality of transmitted audio services by attenuating sound from unwanted sources (e.g. by the application of beamforming techniques). The block diagram of the figure shows one possible implementation of the principles of the present invention. Audio information is acquired via an array of two or more microphones (11, 12) while picture information is acquired via one or more cameras (13) and these signals are then processed as shown in the figure.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

Compression and coding of audio-visual services

Introduction

The present invention relates to the field of audio-visual services over a telecommunications network and in particular, the invention provides techniques for improved signal compression.

Background

Over the past few years, there has been a large growth in the use of telecommunications networks to carry digital audio-visual services, such as video conferencing. The growth of these services is limited by the availability of networks able to provide sufficient capacity to carry the services at acceptable quality and at reasonable cost. In most audiovisual applications, compression techniques are used to reduce the amount of data that must be transmitted and, through this, service cost. The high level of interest in this technology has resulted in several international standards being created to meet different applications (eg. ITU-T H.320 family for video conferencing, MPEG 1 for audio-visual services at approximately 1.5 Mbit/s).

This compression is of a type known as "lossy"; ie. it usually results in a reduction in audio-visual quality. There is a cost-performance trade-off that has led to much research into techniques to maximise the quality achieved for a given level of usage of network resources (usually expressed as the channel capacity in bits per second). One of the key decisions that must be made is to identify those parts of the picture which are most important to a user's perception of quality, and should therefore be transmitted at the highest quality. The remainder of the picture can then be transmitted at a lower quality. Many techniques have been proposed for achieving this, based solely on the use of information contained in the video pictures. These techniques share two common disadvantages: 1) they are very expensive to implement in real time (some are in fact impossible to implement in real time), and 2) they tend to make restrictive assumptions about the video material to be processed (for example, that the picture contains background, plus the head and shoulders of one person).

Summary of the invention

For the purpose of this specification, the term "video quality" is defined as the fidelity with which the dynamic image or part of the dynamic image of a field of view is represented by a video signal, or compressed video

signal, including both the spatial and temporal resolution of the respective image or partial image.

According to a first aspect the present invention consists in an audio-visual signal processing system including:

a) video signal input means for receiving a primary video signal representing an image of a three dimensional space;

b) sound signal input means for receiving a sound signal representing sounds including a sound produced within the three dimensional space;

c) Direction Information Extraction means arranged to process the video and/or sound signals to extract information indicative of a location of a sound source within the three dimensional space; and

d) video signal processing means arranged to identify portions of the primary video signal corresponding to an image encompassing the location of the sound source within the three dimensional space and to produce a secondary video signal from the primary video signal, in which portions corresponding to the location of the sound source within the three dimensional space have a higher video quality (as hereinbefore defined) than at least part of the remainder of the secondary video signal.

According to a second aspect the present invention consists in an audio-visual signal processing system including:

a) video signal input means for receiving a primary video signal representing an image of a three dimensional space;

b) sound signal input means for receiving a sound signal representing sounds including a sound produced within the three dimensional space;

c) Direction Information Extraction means arranged to process the video and/or sound signals to extract information indicative of a location of a sound source of interest within the three dimensional space; and

d) audio processing means arranged to attenuate components of the sound signals representative of sounds not originating from the location of the sound source of interest.

According to a third aspect the present invention consists in a method of processing a video signal in an audio visual system including the steps of:

a) inputting a primary video signal representing an image of a three dimensional space;

b) inputting a sound signal representing sounds including a sound produced within the three dimensional space;

c) processing the primary video signal and/or the sound signal to produce location data representing a location of a sound source within the three dimensional source;

d) using the location data to identify portions of the primary video signal corresponding to an image portion encompassing the location of the sound source; and

e) processing the primary video signal to produce a secondary video signal in which portions of the secondary video signal representing the image portion encompassing the location of the sound source have a higher video quality (as hereinbefore defined) than at least part of the remainder of the secondary video signal.

According to a fourth aspect the present invention consists in a method of processing an audio signal in an audio-visual system including the steps of:

a) inputting a primary video signal representing an image of a three dimensional space directing a camera at a field of view containing a sound source to produce a primary video signal representing a dynamic image of a three dimensional space;

b) inputting a sound signal representing sounds including a sound produced within the three dimensional space;

c) processing the primary video signal and/or the sound signal to produce location data representing a location of a sound source within the three dimensional source;

d) processing the sound signal to selectively attenuate components of the sound signal representative of sounds from different sound sources depending upon the location of each source relative to the sound source represented by the location data.

Preferably, the audio-visual processing system is part of an audio-visual transmission system, including a video signal generating means such

as a video camera or video tape recorder and a sound signal generating means such as a microphone or similar transducer, or a tape recorder or the audio output of a video tape recorder.

In the preferred system, a plurality of audio signals are used, each generated by a microphone at a one of a plurality of different locations around the three dimensional space and in the case of signals retrieved from a tape recorder, the signal from each microphone is recorded on a different recording channel or is multiplexed with other signals in such a way that it can be separated without loss of phase relative to other signals and separately processed.

Preferably, the location of sound sources is achieved by the correlation of the plurality of sound signals, and the locations are mapped into a co-ordinate system from which the images of the location can be identified in the video signal.

Alternatively, the location of sounds sources of interest can be identified by identifying areas of movement within the video image (eg., lip areas) and these image areas can be used to identify sound components issuing from the source of interest by mapping the image areas onto a suitable co-ordinate system and identifying the sounds originating from those co-ordinates.

Other areas of interest, apart from sound sources, may also be identified by identifying areas of movement (eg., eye movement) in which case the secondary video signal may be given higher video quality in those areas as well.

In some embodiments, a plurality of video cameras (or other sources) are used to provide video signals of images from multiple points of view by processing such multiple images simultaneously. By processing such multiple images simultaneously, three dimensional locations of sound sources can be more accurately identified and closer correlation between video images and sound sources can be determined.

Preferably, the video signal is transmitted as a digitally encoded signal in which signals representing those portions of the image which change rapidly are transmitted more frequently than signals representing portions of the image that change less frequently or do not change at all. Preferably, the sound signals are also represented digitally and are multiplexed with the video signals.

According to a fifth aspect, the present invention provides an audio-visual transmitter which includes:

a) video input means to receive video information representing an image of a target;

b) sound input means to receive audio information which permits the location of a source of sound in three dimensional space relative to the target;

c) correlation means to map the audio information onto the image information and to identify a portion of the image information corresponding to an image area encompassing the location of the sound source; and

d) communication means arranged to generate modified video information wherein video quality (as defined herein) of the image represented by the modified video information varies with proximity to the image area encompassing the location of the sound source, the communication means being further arranged to transmit or communicate the modified video information and the audio information to a remote location.

In a preferred embodiment, the video quality of the video information corresponding to the area of the image encompassing the location of the sound source is improved relative to the rest of the video image.

The basis of the new techniques proposed in embodiments of the present invention is the use of directional information in the audio as well as the raw video data to identify regions of interest in pictures so that these regions can be transmitted at higher quality than other less important parts. In addition, cancellation of extraneous audio sources will also be possible, thus enhancing the overall sound quality. These techniques can be applied in many applications of audiovisual compression technology where there is a relationship between the locations of sound sources and important regions in video pictures.

Brief description of the drawings

Embodiments of the invention will now be described in greater detail with reference to the accompanying drawings in which:

Figure 1 is a block diagram of an audio-visual compression system in accordance with the present invention;

Figure 2 depicts a camera and microphone arrangement diagram for an embodiment of the invention;

Figure 3 illustrates the definition of "horizontal direction of arrival"; and

Figure 4 illustrates the definition of "vertical direction of arrival"

Detailed description of the preferred embodiments

In the last decade, audio-visual services have grown beyond their traditional applications in broadcast television to include a large range of interactive services, such as video conferencing. This growth of point to point, as opposed to broadcast, applications has led to demand from users to reduce the cost of transmitting these services. Fortunately, there have been technological advances that permit higher quality services to be delivered using the same resources, or, alternatively, a service of a given quality to be delivered using less resources (and therefore to be delivered more cheaply). One of the significant advances has been the development of the digital approach, in which both the video and audio signals are represented as a sequence of binary numbers.

When audio-visual services are transmitted on digital communications systems, it is necessary to apply compression techniques in order that the amount of information to be transmitted is not beyond the capacity of the network. These techniques are all based on removing the redundant information. For example, video consists of a sequence of frames, each of which is represented by a number of pixels. Each pixel represents the brightness and colour information at a particular point in the frame. However, adjacent pixels tend to be very similar. Also, pixels in the same location in different frames that are closely spaced in time tend be similar. By taking advantage of this 'sameness', it is possible to transmit video information more efficiently. Furthermore, it is not necessary that an exact copy of the video signal be transmitted, since most video contains detail that the human eye cannot see. This observation leads to another method by which compression can be achieved: throwing away information that is not perceptually important.

One of the key limitations in all existing video technology is that the audio and video parts of the information are represented completely independently, in spite of the fact that they are often very strongly related. The present invention provides a first step towards implementing a

combined representation of these two signals. This is widely regarded as a very important problem (for example papers on this topic are regularly solicited for international conferences in this area); it is also regarded as being a very difficult problem. Embodiments of the present invention provide methods whereby audio information can be used to produce an improved representation of the video, more efficiently than this could be achieved by looking at the video alone.

As a simple example of how the audio and video information might be jointly used to improve service quality, we suggest the following. It is well known that not all areas of a video picture are equally important to a viewer's perception of service quality. For example, in a video conference the quality with which the speaker's lips are reproduced is very important, but the quality of the background and other parts of the image is less important. It turns out that to locate accurately the lips of a speaker using the video information alone is a very difficult problem. Solutions proposed so far share two common important disadvantages. Firstly, they are computationally very intensive, and secondly, they do not work well where there is more than one person in the picture. One the other hand, it is relatively simple to use the audio information to identify the location from which a sound emanates (ie. the lips of a person). The audio-based techniques that could be used here are based on ideas of beam-forming. Hence, by combining the audio and video information, it should be possible to achieve a significant increase in the service quality with a relatively small increase in cost.

Embodiments of the invention will be of use in many applications. Two important examples are video conferencing and distance education. In both of these applications, accurate identification of the location of a speaker's lips will enable improved video quality to be achieved within the available capacity of the communications links used to transmit these services.

Most modern video coders use motion estimation to form a motion compensated prediction of the video frame to be encoded. This technique effectively exploits temporal redundancy within the sequence. Next, the Discrete Cosine Transform (DCT) is employed to small blocks of data (usually 8x8 pixels) within the difference image produced by subtracting the motion compensated prediction from the video frame. The DCT both

removes any remaining spatial redundancy in the difference image and also compacts the remaining energy into a relatively few DCT coefficients. The DCT coefficients are then quantised (rounded) and entropy coded (common values are coded with a short codeword while less likely coefficients are coded with longer codewords) prior to transmission to the decoder.

The only part of the coding process which introduces distortion is the quantisation stage since at this point detailed information on DCT coefficient values is irrecoverably discarded. As a result, the quantisation strategy employed plays a crucial part in determining the subjective quality of the received video information. A large amount of research has been carried out in an attempt to quantify what information is important to a viewer. It is well known, for example, that high spatial or temporal activity within a video sequence will mask the effect of coding artefacts. It is also well known that some areas in a sequence are more important than others as far as a viewer is concerned. In the case of a videophone or video conference link, the representation of the mouth of the speaker is of key importance. Unfortunately, attempting to automatically locate the position of the mouth from the video information alone is a difficult problem.

If the position of the mouth could be accurately located then a number of possibilities exist, even within existing internationally standardised video coding approaches, to improve the quality of the delivered service. For example, the mouth area could be more finely quantised thus resulting in a higher quality rendition. Further, most videophone and video conference links utilise a frame rate which is low (say 6 frames per second) compared with the frame rate available from the camera producing the analogue video to the coder (25 frames per second). It would be possible to transmit frames which only contain lip information thus providing a full refresh rate for this important information (and so maintaining good lip sync between video and audio) while only refreshing other areas of the video material at a lower rate. In the remainder of this application, we describe a technique by which this can be achieved.

In embodiments of the present invention, it is proposed that the position of the lips of the speaker in question be known precisely so that different video coding techniques can be used for this area compared to the rest of the scene. It is well known that any deterioration of the reconstruction of the lips destroys the quality of the video interaction more

than any other part of the scene. As the lips of the speaker are the source of sound, it is aimed to determine their positions by estimating the direction of this sound source using an array of microphones suitably placed in a plane near the camera and in front of the speaker.

Direction of arrival estimation techniques using an array of sensors are well known for radar, sonar and seismology. There are many direction finding techniques and all these techniques exploit the fact that time taken by the signal emitted by a source to reach different sensors is different due to the spatial spread of these sensors. The present application requires processing of very wideband signals using a relatively small array for direction of arrival estimation. Processing of broadband array signals is normally carried out by using a tapped delay line filter placed behind each microphone with the signal from each tap being amplified before summing to produce the array output. By solving a constrained beamforming problem one is able determine the optimal gains of these amplifiers such that the array output is optimised to receive an undistorted signal from a point source under consideration and simultaneously reducing the strength of noise arriving from any other direction. This process allows one to measure the power arriving from a point source under consideration.

The present application requires searching for a sound source located at the lips of the speaker. One possible technique for locating this source is to consider points on a spiral starting at the outer boundary of the area under consideration and measure the power arriving from the direction of each point using the optimal beamforming methods described above. The point which yields the maximum power would be identified as the location of the lips.

The microphone array is connected to an appropriate digital signal processor (eg. a member of the Motorola DSP56000 family) which will be used to control the microphone array in real time and thus extract the position of the audio source. The information gathered by the microphone array is utilised by the video coding algorithm to determine the position of the lips of the user.

Another feature of microphone arrays is that they can be used for noise cancellation. Thus once the audio source has been located, a beamforming approach can be used to cancel noise from other sources such as background noise. As well as improving the overall service quality at the

receiver, this approach will also be beneficial in improving the coding efficiency of the audio coding algorithm used.

As mentioned previously, certain parts of a video conferencing picture are more important to a user's perception of visual quality than others. For example, accurate representation of the lips is most important, followed by the eyes and the rest of the face. The remainder of a speaker's body and the scene background are very much less important. The techniques proposed here can be used to quickly and easily locate the lips of each person in a picture who is currently speaking.

This technique could be implemented in conjunction with an existing block-based compression standard (such as ISO/IEC MPEG I and MPEG 2, or ITU-T H.261 and H.263) by choosing a smaller value for the quantisation parameter in those macroblocks containing the most important picture information, and a larger value for the quantisation parameter in other macroblocks.

Two new methods for improving audio visual quality are proposed:

1. The use of directional information obtained from sound (eg. by the use of two or more microphones) to improve the quality of video services by using this directional information to identify important parts of the picture, and transmitting these parts at higher quality (eg. by adjusting quantisation parameters used in the compression process.)
2. The use of directional information obtained from sound and from the video (eg. by picture segmentation techniques) to improve the quality of transmitted audio services by attenuating sound from unwanted sources (eg. by the application of beamforming techniques).

The block diagram of Figure 1 shows one possible implementation of the principles of the present invention. Audio information is acquired via an array of two or more microphones 11, 12 while picture information is acquired via one or more cameras 13 and these signals are them processed as shown in Figure 1.

The Extract Directional Information block 15 identifies the location in the field of view of the camera 13 and microphone array 11, 12 those areas that are of most interest. The direction of sound sources can be identified by measuring the difference in propagation delay between the sound source and

two or more microphones 11, 12. This can be performed by examining the cross correlation between the signals from the different microphones 11 and 12. Where the delay difference is small, it may be possible to use a subtraction operation rather than multiplication in the cross correlation. The video information can be used by looking for regions with high motion. An appropriate transformation of coordinate systems is required so that the results obtained from audio and video information are specified with respect to the same coordinate system.

In the Preprocess Audio block 14 the directional information is used to apply beamforming techniques to select particular directions or to attenuate particular regions of the field of view of the microphone array. The simplest action for this block would be to either select a single microphone or sum the outputs of all microphones together. For compression schemes that support multiple channels, the output could be multi-channel audio.

The Video Compression block 16 uses a video compression algorithm, which takes in video pictures and produces a compressed video bitstream, eg. ISO/IEC MPEG 2 Video, ITU-T H.263. This block may include filtering to reduce the spatial and/or temporal resolution, or for other purposes

The Audio Compression block 17 uses an audio compression algorithm, which takes in a digital audio stream, and produces a compressed audio bitstream. Examples include ISO/IEC MEG 2 Audio, & ITU T G.723.

The multiplexer 18 combines the various video and audio bitstreams so that they can be carried on a single channel, eg. ITU-T H.922 and ISO/IEC MPEG 2 Systems.

Preferred embodiments of the invention employing these new techniques

1. cope easily with moving objects in a picture,
2. deal with any number of important objects present in a picture,
3. have very much lower cost for real-time implementation,
4. can be implemented in ways that are compatible with existing international standards for audio-visual compression,
5. do not require any operator intervention.

Referring now to Figures 2, 3, and 4 the application of one embodiment of the invention can be achieved by the following steps:

1. suitable arrangement of microphones 11, 12, 21 and camera 13,

2. use of sound arriving at the various microphones to identify the direction of origin of the sound,

3. translation of this direction of origin of the sound source to the corresponding location in the video picture,

4. use of the information on sound source location in the video picture to improve the quality of the video service.

As illustrated in Figure 2, three microphones 11, 12, 21 and a camera 13 are arranged with the microphones 11, 12, 21 placed at the corners of a square 22 that has sides of length d_m. In this implementation, d_m=1m. The camera 13 is placed at the centre of the square 22, with its imaging plane lying in the plane of the microphones 11, 12, 21 and the centre of its field of view perpendicular to this plane. Each frame in the camera output consists of v lines, with each line consisting of h pixels.

The meaning of the horizontal angle of arrival θ_H of the sound is shown in Figure 3 under the assumption that the sound source is not too close to the microphones 11, 12. The arrow indicates the direction of travel of the sound. $\theta_H>0$ indicates that the sound source is to the right of the perpendicular to the plane of the microphones; $\theta_H<0$ indicates that the sound source is to the left of the perpendicular.

The meaning of the vertical angle of arrival θ_V of the sound is shown in Figure 5 under the assumption that the sound source is not too close to the microphones 11, 21. The arrow indicates the direction of travel of the sound $\theta_V>0$ indicates that the sound source is above the perpendicular to the plane of the microphones; $\theta_H<0$ indicates that the sound source is below the perpendicular.

The field of view of the camera 13 is the area between an angle θ_{Hmax} to the left and right of the perpendicular to the plane of the microphones and an angle θ_{Vmax} above and below this perpendicular. Hence, the angle of the field of view is $2\theta_{Hmax}$ horizontally and $2\theta_{Vmax}$ vertically.

The output of each microphone is connected to a 16 bit analog to digital converter, which has sampling frequency f_s, which here takes the value of 32,000 samples per second. The sequence of samples output from the analog to digital converter connected to the first microphone 11 is denoted by $x_1(t)$, where t is the time at which the sample is taken. Similarly, the sequence of samples from the second and third microphones 12, 21, are

denoted by $x_2(t)$ and $x_3(t)$ respectively. Assuming that the sound sources are not too close to the microphones and that the sound sources of interest lie within the field of view of the camera, the horizontal angle of arrival θ_H is calculated for each video frame by:

$$\theta_H = \sin^{-1}\left(\frac{v_s \arg\max_{-iH<i<iH} \sum_{j=1}^{2000} x_1(k+j)x(k+j-i)}{d_m f_s} \right)$$

where k denotes the time corresponding to the beginning of the video frame and iH is the value of i that corresponds to a sound source whose horizontal location places it on the edge of the field of view of the camera (i.e. $\theta_H = \theta_{Hmax}$).

The vertical angle of arrival θ_V for each video frame is given by:

$$\theta_V = \sin^{-1}\left(\frac{v_s \arg\max_{-iV<i<iV} \sum_{j=1}^{2000} x_1(k+j)x_3(k+j-i)}{d_m f_s} \right)$$

where k denotes the time corresponding to the beginning of the video frame and iV is the value of i that corresponds to a sound source whose vertical location places it on the edge of the field of view of the camera (i.e. $\theta_V = \theta_{Vmax}$).

After calculation of θ_H and θ_V, the location of the sound source in the video pictures can be calculated. This location lies to the right of the center of the picture by an amount:

$$\frac{h \tan\theta_H}{2 \tan\theta_{H\max}} \ pixels$$

and above the center of the picture by

$$\frac{v\tan\theta_v}{2\tan\theta_{vmax}}\ pixels$$

The information on the location of sound sources is applied to a video coding system conforming with the ITU-T- Recommendation H.261.

Macroblocks each picture belong to one of two classes:

1. Macroblocks containing the block containing the location of the sound source (as calculated in the previous section) or any block immediately adjacent to this macroblock There will be four macroblocks in this class for each picture.
2. All other macroblocks.

The information on the location of sound sources is applied to the video coding by using two different schemes for choosing the value of the H.261 parameter QP for macroblocks in the two classes.

In a first implementation a variable bitrate is employed. In this implementation, for macroblocks in class I, we choose QP = 4 and for macroblocks in class 2, we choose QP = 16.

In a second implementation a constant bitrate is employed. In this implementation for macroblocks in class 1, we choose QP = 4 and for macroblocks in class 2, we choose QP according to the rule described in version 8 of the Test Model used in developing the ITU-T H.263 Recommendation

It will be appreciated by persons skilled in the art that numerous variations and/or modifications may be made to the invention as shown in the specific embodiments without departing from the spirit or scope of the invention as broadly described. The present embodiments are, therefore, to be considered in all respects as illustrative and not restrictive.

CLAIMS:

1. An audio-visual signal processing system including:

a) video signal input means for receiving a primary video signal representing an image of a three dimensional space;

b) sound signal input means for receiving a sound signal representing sounds including a sound produced within the three dimensional space;

c) Direction Information Extraction means arranged to process the video and/or sound signals to extract information indicative of a location of a sound source within the three dimensional space; and

d) video signal processing means arranged to identify portions of the primary video signal corresponding to an image encompassing the location of the sound source within the three dimensional space and to produce a secondary video signal from the primary video signal, in which portions corresponding to the location of the sound source within the three dimensional space have a higher video quality (as hereinbefore defined) than at least part of the remainder of the secondary video signal.

2. The signal processing system as claimed in claim 1, wherein a plurality of sound input means are provided, each arranged to accept a different sound input, each sound input being originally generated by a sound transducer at one of a plurality of different locations in or around the three dimensional space.

3. The signal processing system as claimed in claim 1 or 2, wherein a plurality of video signal input means are provided, each arranged to receive a different video signal.

4. The signal processing system as claimed in claim 3, wherein each video signal input means receives a video signal from a video camera directed at the three dimensional space.

5. The signal processing system as claimed in claims 1, 2, 3, or 4, wherein each sound transducer is connected to the signal processing system.

6. The signal processing system as claimed in claim 1, 2, 3 or 4, wherein each video signal input means receives a video signal from a video recording system, each video signal being originally recorded from a video camera directed at the three dimensional space.

7. The signal processing system as claimed in claim 6, wherein the plurality of sound inputs are provided by a multichannel recording system each sound input being recorded on a separate channel of the recording system, and each channel being originally recorded from a transducer located at one of the plurality of different locations in or around the three dimensional space.

8. The signal processing system as claimed in claim 7, wherein the sound recording system is incorporated into the video recording, whereby the sound and video signals are synchronised.

9. The signal processing system as claimed in claim 7, whereby the sound recording system and video recording system are independent systems interconnected by synchronising means.

10. The signal processing system as claimed in any one of claims 1 to 9. wherein the direction information extraction means, includes correlation means arranged to correlate the plurality of sound signals with one another to produce location information mapping sound sources within the three dimensional space.

11. The signal processing system as claimed in any one of claims 1 to 10, wherein the direction extraction information means includes video signal processing means arranged to identify areas of movement within the three dimensional space and to produce location information mapping the movement locations within the three dimensional space.

12. The signal processing system as claimed in any one of claims 1 to 11, wherein the direction information extraction means includes correlation means arranged to correlate the plurality of video signals with one another to produce three dimensional location information mapping areas of motion or other areas of interest within the three dimensional space.

13. The signal processing system as claimed in any one of claims 10, 11 or 12, further including correlation means arranged to correlate the sound source location information and the location information relating to areas of motion or other areas of interest within the three dimensional space whereby particular sound components are associated with portions of the video image with which their source location corresponds.

14. The signal processing system as claimed in any one of claims 1 to 13, wherein the secondary video signal is a digital video signal and the portion

of the video signal corresponding to the sound source is generated more frequently than the remainder of the video signal.

15. The signal processing system as claimed in any one of claims 1 to 14, wherein a component of the sound signal associated with an area of motion in the three dimensional space is enhanced relative to the other components of the sound signal.

16. An audio-visual signal processing system including:

a) video signal input means for receiving a primary video signal representing an image of a three dimensional space;

b) sound signal input means for receiving a sound signal representing sounds including a sound produced within the three dimensional space;

c) Direction Information Extraction means arranged to process the video and/or sound signals to extract information indicative of a location of a sound source of interest within the three dimensional space; and

d) audio processing means arranged to attenuate components of the sound signals representative of sounds not originating from the location of the sound source of interest.

17. The signal processing system as claimed in claim 16, wherein a plurality of sound input means are provided, each arranged to accept a different sound input, each sound input being originally generated by a sound transducer at one of a plurality of different locations in or around the three dimensional space.

18. The signal processing system as claimed in claim 16 or 17, wherein a plurality of video signal input means are provided, each arranged to receive a different video signal.

19. The signal processing system as claimed in claim 18, wherein each video signal input means receives a video signal from a video camera directed at the three dimensional space.

20. The signal processing system as claimed in claims 16, 17, 18 or 19, wherein each sound transducer is connected to the signal processing system.

21. The signal processing system as claimed in claims 16, 17, 18, or 19, wherein each video signal input means receives a video signal from a video recording system, each video signal being originally recorded from a video camera directed at the three dimensional space.

22. The signal processing system as claimed in claim 21, wherein the plurality of sound inputs are provided by a multichannel recording system each sound input being recorded on a separate channel of the recording system, and each channel being originally recorded from a transducer located at one of the plurality of different locations in or around the three dimensional spaces.

23. The signal processing system as claimed in claim 22, wherein the sound recording system is incorporated into the video recording, whereby the sound and video signals are synchronised.

24. The signal processing system as claimed in claim 22, whereby the sound recording system and video recording system are independent systems interconnected by synchronising means.

25. The signal processing system as claimed in any one of claims 16 to 24, wherein the direction information extraction means, includes correlation means arranged to correlate the plurality of sound signals with one another to produce location information mapping sound sources within the three dimensional space.

26. The signal processing system as claimed in any one of claims 16 to 25, wherein the direction extraction information means includes video signal processing means arranged to identify areas of movement within the three dimensional space and to produce location information mapping the movement locations within the three dimensional space.

27. The signal processing system as claimed in any one of claims 16 to 26, wherein the direction information extraction means includes correlation means arranged to correlate the plurality of video signals with one another to produce three dimensional location information mapping areas of motion or other areas of interest within the three dimensional space.

28. The signal processing system as claimed in any one of claims 25 to 27, further including correlation means arranged to correlate the sound source location information and the location information relating to areas of motion or other areas of interest within the three dimensional space whereby particular sound components are associated with portions of the video image with which their source location corresponds.

29. The signal processing system as claimed in any one of claims 16 to 28, wherein a portion of the secondary video signal corresponding to a sound

source location is enhanced relative to the other components of the video signal.

30. The signal processing system as claimed in any one of claims 16 to 29 , wherein the secondary video signal is a digital video signal and the portion of the video signal corresponding to the sound source is generated more frequently than the remainder of the video signal.

31. An audio-visual transmitter which includes:

a) video input means to receive video information representing an image of a target;

b) sound input means to receive audio information which permits the location of a source of sound in three dimensional space relative to the target;

c) correlation means to map the audio information onto the image information and to identify a portion of the image information corresponding to an image area encompassing the location of the sound source; and

d) communication means arranged to generate modified video information wherein video quality (as defined herein) of the image represented by the modified video information varies with proximity to the image area encompassing the location of the sound source, the communication means being further arranged to transmit or communicate the modified video information and the audio information to a remote location.

32. The transmitter as claimed in claim 31, wherein the video quality of the modified video information increases with proximity to the image area encompassing the location of the sound source.

33. The transmitter as claimed in claim 31 or 32, wherein a plurality of sound input means are provided, each arranged to accept a different sound input, each sound input being originally generated by a sound transducer at one of a plurality of different locations in or around the three dimensional space.

34. The transmitter as claimed in claim 31, 32 or 33, wherein a plurality of video signal input means are provided, each arranged to receive a different video signal.

35. The transmitter as claimed in claim 34, wherein each video signal input means receives a video signal from a video camera directed at the three dimensional space.

36. The transmitter as claimed in claim 31, 32, 33, 34 or 35, wherein the sound transducer is connected to the signal processing system.

37. The transmitter as claimed in any one of claims 31 to 35, wherein each video signal input means receives a video signal from a video recording system, each video signal being originally recorded from a video camera directed at the three dimensional space.

38. The transmitter as claimed in claim 37, wherein the plurality of sound inputs are provided by a multichannel recording system each sound input being recorded on a separate channel of the recording system, and each channel being originally recorded from a transducer located at one of the plurality of different locations in or around the three dimensional spaces.

39. The transmitter as claimed in claim 38, wherein the sound recording system is incorporated into the video recording, whereby the sound and video signals are synchronised.

40. The transmitter as claimed in claim 38, whereby the sound recording system and video recording system are independent systems interconnected by synchronising means.

41. The transmitter as claimed in any one of claims 31 to 40, wherein the direction information extraction means, includes correlation means arranged to correlate the plurality of sound signals with one another to produce location information mapping sound sources within the three dimensional space.

42. The transmitter as claimed in any one of claims 31 to 41, wherein the direction extraction information means includes video signal processing means arranged to identify areas of movement within the three dimensional space and to produce location information mapping the movement locations within the three dimensional space.

43. The transmitter as claimed in any one of claims 31 to 42, wherein the direction information extraction means includes correlation means arranged to correlate the plurality of video signals with one another to produce three dimensional location information mapping areas of motion or other areas of interest within the three dimensional space.

44. The transmitter as claimed in claim 41, 42 or 43, further including correlation means arranged to correlate the sound source location information and the location information relating to areas of motion or other areas of interest within the three dimensional space whereby particular sound components are associated with portions of the video image with which their source location corresponds.

45. The transmitter as claimed in any one of claims 31 to 44, wherein the secondary video signal is a digital video signal and the portion of the video signal corresponding to the sound source is generated more frequently than the remainder of the video signal.

46. The transmitter as claimed in any one of claims 31 to 45, wherein a component of the sound signal associated with an area of motion in the three dimensional space is enhanced relative to the other components of the sound signal.

47. A method of processing a video signal in an audio visual system including the steps of:

a) inputting a primary video signal representing an image of a three dimensional space;

b) inputting a sound signal representing sounds including a sound produced within the three dimensional space;

c) processing the primary video signal and/or the sound signal to produce location data representing a location of a sound source within the three dimensional source;

d) using the location data to identify portions of the primary video signal corresponding to an image portion encompassing the location of the sound source; and

e) processing the primary video signal to produce a secondary video signal in which portions of the secondary video signal representing the image portion encompassing the location of the sound source have a higher video quality (as hereinbefore defined) than at least part of the remainder of the secondary video signal.

48. The method as claimed in claim 47, wherein a sound signal is input to each of a plurality of sound input means, each arranged to accept a different sound input and each sound input originally generated by a sound transducer at one of a plurality of different locations in or around the three dimensional space.

49. The method as claimed in claim 47 or 48, wherein a video signal is input to each of a plurality of video signal input means, each arranged to receive a different video signal.

50. The method as claimed in claim 49, wherein each video signal input means receives a video signal from a video camera directed at the three dimensional space.

51. The method as claimed in any one of claims 47 to 50, wherein each sound transducer is connected to the signal processing system.

52. The method as claimed in any one of claims 47 to 49, wherein each video signal input means receives a video signal from a video recording system, each video signal being originally recorded from a video camera directed at the three dimensional space.

53. The method as claimed in claim 52, wherein the plurality of sound inputs are provided by a multichannel recording system each sound input being recorded on a separate channel of the recording system, and each channel being originally recorded from a transducer located at one of the plurality of different locations in or around the three dimensional spaces.

54. The method as claimed in claim 53, wherein the sound recording system is incorporated into the video recording, whereby the sound and video signals are synchronised.

55. The method as claimed in claim 53, whereby the sound recording system and video recording system are independent systems interconnected by synchronising means.

56. The method as claimed in any one of claims 47 to 55, wherein the location data includes data produced by correlating the plurality of sound signals with one another such that the location data maps sound sources within the three dimensional space.

57. The method as claimed in any one of claims 47 to 56, wherein the location data includes data produced by identifying areas of movement within the three dimensional space such that the location data maps the movement locations within the three dimensional space.

58. The method as claimed in any one of claims 47 to 57, wherein the location data includes data produced by correlating the plurality of video signals with one another to produce three dimensional location data mapping areas of motion or other areas of interest within the three dimensional space.

59. The method as claimed in claim 56, 57 or 58, wherein the location data further includes data produced by correlating the sound source location information and the location information relating to areas of motion or other areas of interest within the three dimensional space whereby particular sound components are associated with portions of the video image with which their source location corresponds.

60. The method as claimed in any one of claims 47 to 59, wherein the secondary video signal is a digital video signal and the portion of the video signal corresponding to the sound source is generated more frequently than the remainder of the video signal.

61. The method as claimed in claim 60, wherein a component of the sound signal associated with an area of motion in the three dimensional space is enhanced relative to the other components of the sound signal.

62. A method of processing an audio signal in an audio-visual system including the steps of:

a) inputting a primary video signal representing an image of a three dimensional space directing a camera at a field of view containing a sound source to produce a primary video signal representing a dynamic image of a three dimensional space;

b) inputting a sound signal representing sounds including a sound produced within the three dimensional space;

c) processing the primary video signal and/or the sound signal to produce location data representing a location of a sound source within the three dimensional source;

d) processing the sound signal to selectively attenuate components of the sound signal representative of sounds from different sound sources depending upon the location of each source relative to the sound source represented by the location data.

63. The method as claimed in claim 62, wherein a sound signal is input to each of a plurality of sound input means are provided, each arranged to accept a different sound input and each sound input originally generated by a sound transducer at one of a plurality of different locations in or around the three dimensional space.

64. The method as claimed in claim 62 or 63, wherein a video signal is input to each of a plurality of video signal input means, each arranged to receive a different video signal.

65. The method as claimed in claim 64, wherein each video signal input means receives a video signal from a video camera directed at the three dimensional space.

66. The method as claimed in any one of claims 62 to 65, wherein each sound transducer is connected to the signal processing system.

67. The method as claimed in any one of claims 62 to 64, wherein each video signal input means receives a video signal from a video recording system, each video signal being originally recorded from a video camera directed at the three dimensional space.

68. The method as claimed in claim 67, wherein the plurality of sound inputs are provided by a multichannel recording system each sound input being recorded on a separate channel of the recording system, and each channel being originally recorded from a transducer located at one of the plurality of different locations in or around the three dimensional space.

69. The method as claimed in claim 68, wherein the sound recording system is incorporated into the video recording, whereby the sound and video signals are synchronised.

70. The method as claimed in claim 68, whereby the sound recording system and video recording system are independent systems interconnected by synchronising means.

71. The method as claimed in any one of claims 62 to 70, wherein the location data includes data produced by correlating the plurality of sound signals with one another such that the location data maps sound sources within the three dimensional space.

72. The method as claimed in any one of claims 62 to 71, wherein location data includes data produced by identifying areas of movement within the three dimensional space such that the location data maps the movement locations within the three dimensional space.

73. The method as claimed in any one of claims 62 to 72, wherein the location data includes data produced by correlating the plurality of video signals with one another to produce three dimensional location data mapping areas of motion or other areas of interest within the three dimensional space.

74. The method as claimed in claim 71, 72 or 73, wherein the location data further includes data produced by correlating the sound source location information and the location information relating to areas of motion or other

areas of interest within the three dimensional space whereby particular sound components are associated with portions of the video image with which their source location corresponds.

75. The method as claimed in any one of claims 62 to 74, wherein a portion of the secondary video signal corresponding to a sound source location is enhanced relative to the remainder of the secondary video signal.

76. The method as claimed in any one of claims 62 to 75 , wherein the secondary video signal is a digital video signal and the portion of the video signal corresponding to the sound source is generated more frequently than the remainder of the video signal.

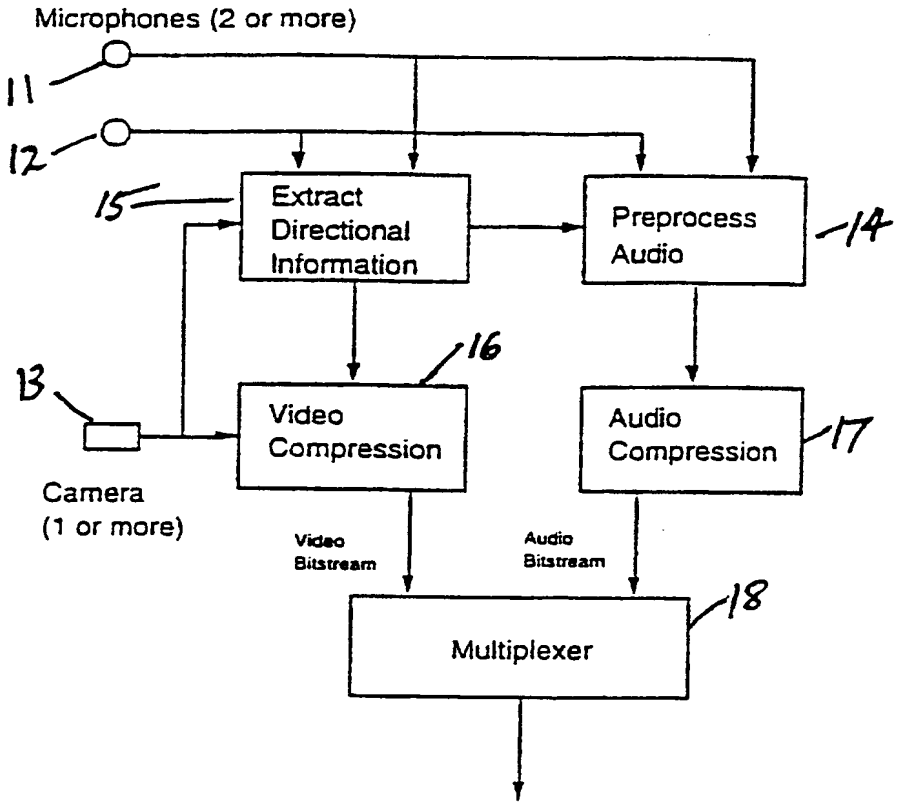


FIGURE 1.

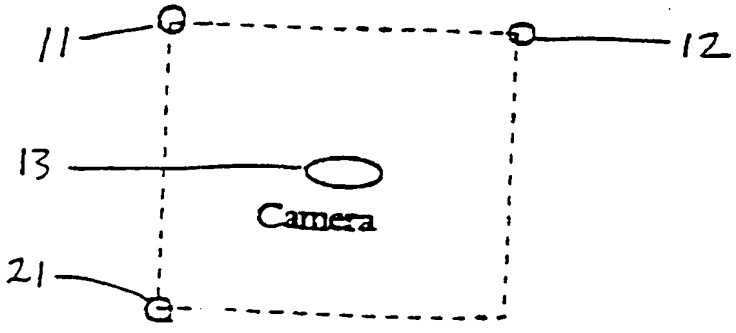


FIGURE 2.

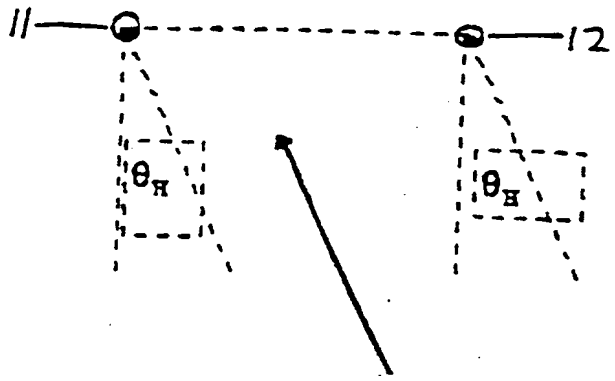


FIGURE 3.

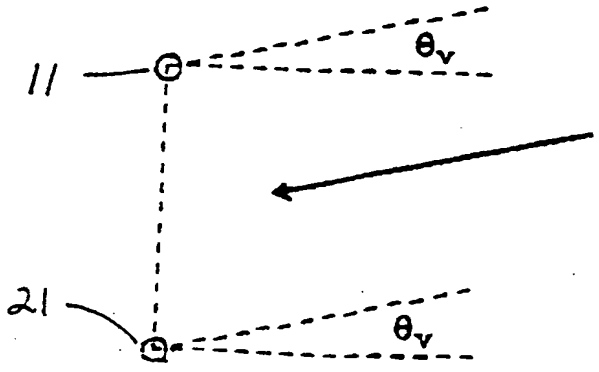


FIGURE 4.

INTERNATIONAL SEARCH REPORT

International Application No.
PCT/AU 97/00297

A. CLASSIFICATION OF SUBJECT MATTER

Int Cl[6]: H04N 7/15; 7/26

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC : H04N 7/15; 7/26

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
AU : IPC as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
DERWENT : COMPRESS
COMPENDEX : VIDEO CONFERENCE; COMPRESSION; TRANSMISSION

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	EP 617537 A1 (AT & T CORP) 28 September 1994 Figure 2	1-76
A	WO 94/00951 A1 (BRITISH TELECOMMUNICATIONS PLC) 6 January 1994 page 6 line 21 - page 7 line 14; figure 2	1-76

☐ Further documents are listed in the continuation of Box C

☒ See patent family annex

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search
19 June 1997

Date of mailing of the international search report
23 JUN 1997

Name and mailing address of the ISA/AU
AUSTRALIAN INDUSTRIAL PROPERTY ORGANISATION
PO BOX 200
WODEN ACT 2606
AUSTRALIA Facsimile No.: (06) 285 3929

Authorized officer

R.G. TOLHURST

Telephone No.: (06) 283 2187

Form PCT/ISA/210 (second sheet) (July 1992) cophin

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No.
PCT/AU 97/00297

This Annex lists the known "A" publication level patent family members relating to the patent documents cited in the above-mentioned international search report. The Australian Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

Patent Document Cited in Search Report		Patent Family Member					
EP	617537 A1	CA	2114868	JP	7095300	US	5390177
WO	94/00951 A1	AU	43517/93	GB	2283636	EP	648400

END OF ANNEX